Electronic Journal of Statistics
Vol. 5 (2011) 192–203
ISSN: 1935-7524
DOI: 10.1214/11-EJS605

A local maximal inequality under uniform entropy

Aad van der Vaart

Department of Mathematics, Faculty of Sciences, Vrije Universiteit De Boelelaan 1081a, 1081 HV Amsterdam,
e-mail: aad@cs.vu.nl

and

Jon A. Wellner*

Department of Statistics, University of Washington, Seattle, WA 98195-4322,
e-mail: jaw@stat.washington.edu

Abstract: We derive an upper bound for the mean of the supremum of the empirical process indexed by a class of functions that are known to have variance bounded by a small constant δ. The bound is expressed in the uniform entropy integral of the class at δ. The bound yields a rate of convergence of minimum contrast estimators when applied to the modulus of continuity of the contrast functions.

AMS 2000 subject classifications: 60K35.
Keywords and phrases: Empirical process, modulus of continuity, minimum contrast estimator, rate of convergence.

Received December 2010.

1. Introduction

The *empirical measure* $\mathbb{P}_n$ and *empirical process* $\mathbb{G}_n$ of a sample of observations $X_1, \ldots, X_n$ from a probability measure P on a measurable space $(\mathcal{X}, \mathcal{A})$ attach to a given measurable function $f: \mathcal{X} \to \mathbb{R}$ the numbers

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad \mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \big(f(X_i) - Pf\big).$$

It is often useful to study the suprema of these stochastic processes over a given class $\mathcal{F}$ of measurable functions. The distribution of the supremum

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$$

is known to concentrate near its mean value, at a rate depending on the size of the envelope function of the class $\mathcal{F}$, but irrespective of its complexity. On the

*Supported in part by NSF Grant DMS-0804587, and by NI-AID grant 2R01 AI291968-04.

other hand, the mean value of $\|\mathbb{G}_n\|_{\mathcal{F}}$ depends on the size of the class $\mathcal{F}$. Entropy integrals, of which there are two basic versions, are useful tools to bound this mean value.

The *uniform entropy integral* was introduced in [9] and [5], following [3], in their study of the abstract version of Donsker's theorem. We define an L_r-version of it as

$$J(\delta, \mathcal{F}, L_r) = \sup_Q \int_0^\delta \sqrt{1 + \log N\bigl(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\bigr)}\, d\varepsilon.$$

Here the supremum is taken over all finitely discrete probability distributions Q on $(\mathcal{X}, \mathcal{A})$, the *covering number* $N\bigl(\varepsilon, \mathcal{F}, L_r(Q)\bigr)$ is the minimal number of balls of radius ε in $L_r(Q)$ needed to cover $\mathcal{F}$, F is an envelope function of $\mathcal{F}$, and $\|f\|_{Q,r}$ denotes the norm of a function f in $L_r(Q)$. The integral is defined relative to an *envelope function*, which need not be the minimal one, but can be any measurable function $F\colon \mathcal{X} \to \mathbb{R}$ such that $|f| \le F$ for every $f \in \mathcal{F}$. If multiple envelope functions are under consideration, then we write $J(\delta, \mathcal{F}|\, F, L_r)$ to stress this dependence. An inequality, due to Pollard (also see [12], 2.14.1), says, under some measurability assumptions, that

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J(1, \mathcal{F}, L_2)\, \|F\|_{P,2}. \tag{1.1}$$

Here $\lesssim$ means smaller than up to a universal constant. This shows that for a class $\mathcal{F}$ with finite uniform entropy integral, the supremum $\|\mathbb{G}_n\|_{\mathcal{F}}$ is not essentially bigger than a multiple of the empirical process $\mathbb{G}_n F$ at the envelope function F. The inequality is particularly useful if this envelope function is small.

The *bracketing entropy integral* has its roots in the Donsker theorem of [8], again following initial work by Dudley. For a given norm it can be defined as

$$J_{[\,]}\bigl(\delta, \mathcal{F}, \|\cdot\|\bigr) = \int_0^\delta \sqrt{1 + \log N_{[\,]}\bigl(\varepsilon \|F\|, \mathcal{F}, \|\cdot\|\bigr)}\, d\varepsilon.$$

Here the *bracketing number* $N_{[\,]}\bigl(\varepsilon, \mathcal{F}, \|\cdot\|\bigr)$ is the minimal number of brackets $[l, u] = \{f\colon \mathcal{X} \to \mathbb{R}\colon l \le f \le u\}$ of size $\|u - l\|$ smaller than ε needed to cover $\mathcal{F}$. A useful inequality, due to Pollard (also see [12], 2.14.2), is

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[\,]}\bigl(1, \mathcal{F}, L_2(P)\bigr)\, \|F\|_{P,2}. \tag{1.2}$$

Bracketing numbers are bigger than covering numbers (at twice the size), and hence the bracketing integral is bigger than a multiple of the corresponding entropy integral. However, the bracketing integral involves only the single distribution P, whereas the uniform entropy integral takes a supremum over all (discrete) distributions, making the two integrals incomparable in general. Apart from this difference the two maximal inequalities have the same message.

The two inequalities (1.1) and (1.2) involve the size of the envelope function, but not the sizes of the individual functions in the class $\mathcal{F}$. They also exploit finiteness of the entropy integrals only, roughly requiring that the entropy grows

at smaller order than ε^{-2} as $\varepsilon \downarrow 0$, and not the precise size of the entropy. In the case of the bracketing integral this is remedied in the equality (see [12], 3.4.2), valid for any class of functions $f: \mathcal{X} \to [-1,1]$ with $Pf^2 \leq \delta^2 PF^2$ and any $\delta \in (0,1)$,

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) \, \|F\|_{P,2} \left(1 + \frac{J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)}{\delta^2 \sqrt{n} \|F\|_{P,2}}\right). \tag{1.3}$$

Here the assumption that the class of functions is uniformly bounded is too restrictive for some applications, but can be removed if the entropy integral is computed relative to the stronger "norm"

$$\|f\|_{P,B} = \Big(2P\big(e^{|f|} - 1 - |f|\big)\Big)^{1/2}.$$

Although it is not a norm, this quantity can be used to define the size of brackets and hence bracketing numbers. Inequality (1.3) is valid for an arbitrary class of functions with $\|f\|_{P,B} \leq \delta \|F\|_{P,B}$ if the $L_2(P)$-norm is replaced by $\|\cdot\|_{P,B}$ in its right side (at four appearances) (see Theorem 3.4.3 of [12]). The "norm" $\|\cdot\|_{P,B}$ derives from the refined version of Bernstein's inequality, which was first used in the literature on rates of convergence of minimum contrast estimators in [1] (also see [11]).

Maximal inequalities of type (1.3) using *uniform entropy* are thus far unavailable. In this note we derive an exact parallel of (1.3) for uniformly bounded functions, and investigate similar inequalities for unbounded functions. The validity of these results seems unexpected, as the stronger control given by bracketing has often been thought necessary for estimates of moduli of continuity. It was suggested to us by Theorem 3.1 and its proof in [4].

1.1. Application to minimum contrast estimators

Inequalities involving the sizes of the functions f are of particular interest in the investigation of empirical minimum contrast estimators. Suppose that $\hat{\theta}_n$ mimimizes a criterion of the type

$$\theta \mapsto \mathbb{P}_n m_\theta,$$

for given measurable functions $m_\theta: \mathcal{X} \to \mathbb{R}$ indexed by a parameter θ, and that the population contrast satisfies, for a "true" parameter θ_0 and some metric d on the parameter set,

$$Pm_\theta - Pm_{\theta_0} \gtrsim d^2(\theta, \theta_0).$$

A bound on the rate of convergence of $\hat{\theta}_n$ to θ_0 can then be derived from the modulus of continuity of the empirical process $\mathbb{G}_n m_\theta$ indexed by the functions m_θ. Specifically (see e.g. [12], 3.2.5) if ϕ_n is a function such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and

$$\mathrm{E} \sup_{\theta: d(\theta,\theta_0) < \delta} \big|\mathbb{G}_n(m_\theta - m_{\theta_0})\big| \lesssim \phi_n(\delta), \tag{1.4}$$

then $d(\hat{\theta}_n, \theta_0) = O_P(\delta_n)$, for δ_n any solution to

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2. \tag{1.5}$$

Inequality (1.4) involves the empirical process indexed by the class of functions $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}$. If d dominates the $L_2(P)$-norm, or another norm $\|\cdot\|$ that can be used in an equality of the type (1.3), such as the Bernstein norm, and the norms of the envelopes of the classes M_δ are bounded in δ, then we can choose

$$\phi_n(\delta) = J\big(\delta, \mathcal{M}_\delta, \|\cdot\|\big)\left(1 + \frac{J\big(\delta, \mathcal{M}_\delta, \|\cdot\|\big)}{\delta^2\sqrt{n}}\right),$$

where J is an appropriate entropy integral. For this choice the inequality (1.5) is equivalent to

$$J\big(\delta_n, \mathcal{M}_{\delta_n}, \|\cdot\|\big) \leq \sqrt{n}\delta_n^2. \tag{1.6}$$

Thus a rate of convergence can be read off directly from the entropy integral.

We note that an inequality of type (1.3) is unattractive for very small δ, as the bound may even increase to infinity as $\delta \downarrow 0$. However, it is accurate for the range of δ that are important in the application to moduli of continuity.

Moduli of continuity also play an important role in model selection theorems. See for instance [7].

Inequalities involving uniform entropy permit for instance the immediate derivation of rates of convergence for minimum contrast functions that form VC-classes. Furthermore, uniform entropy is preserved under various (combinatorial) operations to make new classes of functions. This makes uniform entropy integrals a useful tool in situations where bracketing numbers may be difficult to handle. Equation (1.6) gives an elegant characterization of rates of convergence in these situations, where thus far ad-hoc arguments were necessary.

2. Uniformly bounded classes

Call the class $\mathcal{F}$ of functions *P-measurable* if the map

$$(X_1, \ldots, X_n) \mapsto \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n e_i f(X_i)\right|$$

on the completion of the probability space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ is measurable, for every sequence $e_1, e_2, \ldots, e_n \in \{-1, 1\}$.

Theorem 2.1. *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope function $F \leq 1$ and such that $\mathcal{F}^2$ is P-measurable. If $Pf^2 < \delta^2 PF^2$, for every f and some $\delta \in (0, 1)$, then*

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_\mathcal{F} \lesssim J\big(\delta, \mathcal{F}, L_2\big)\left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2\sqrt{n}\|F\|_{P,2}}\right)\|F\|_{P,2}.$$

Proof. We use the following refinement of (1.1) (see e.g. [12], 2.14.1): for any P-measurable class $\mathcal{F}$,

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \mathrm{E}_P^* J\left(\frac{\sup_f(\mathbb{P}_n f^2)^{1/2}}{(\mathbb{P}_n F^2)^{1/2}}, \mathcal{F}, L_2\right)(\mathbb{P}_n F^2)^{1/2}. \tag{2.1}$$

Because $\delta \mapsto J(\delta, \mathcal{F}, L_2)$ is the integral of a nonincreasing nonnegative function, it is a concave function such that the map $t \mapsto J(t)/t$, which is the average of its derivative over $[0,t]$, is nonincreasing. The concavity shows that its *perspective* $(x,t) \mapsto tJ(x/t, \mathcal{F}, L_2)$ is a concave function of its two arguments (cf. [2], page 89). Furthermore, the "extended-value extension" of this function (which by definition is $-\infty$ if $x \le 0$ or $t \le 0$) is obviously nondecreasing in its first argument and was noted to be nondecreasing in its second argument. Therefore, by the vector composition rules for concave functions ([2], pages 83–87, especially lines -2 and -1 of page 86), the function $(x,y) \mapsto H(x,y) := J\big(\sqrt{x/y}, \mathcal{F}, L_2\big)\sqrt{y}$ is concave. We have that $\mathrm{E}_P^*\mathbb{P}_n F^2 = \|F\|_{P,2}^2$. Therefore, by an application of Jensen's inequality to the right side of the preceding display we obtain, for $\sigma_n^2 = \sup_f \mathbb{P}_n f^2$,

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\left(\frac{\sqrt{\mathrm{E}_P^*\sigma_n^2}}{\|F\|_{P,2}}, \mathcal{F}, L_2\right)\|F\|_{P,2}. \tag{2.2}$$

The application of Jensen's inequality with outer expectations can be justified here by the monotonicity of the function H, which shows that the measurable majorant of a variable $H(U,V)$ is bounded above by $H(U^*, V^*)$, for U^* and V^* measurable majorants of U and V. Thus $\mathrm{E}^* H(U,V) \le \mathrm{E}H(U^*, V^*)$, after which Jensen's inequality can be applied in its usual (measurable) form.

The second step of the proof is to bound $\mathrm{E}_P^*\sigma_n^2$. Because $\mathbb{P}_n f^2 = Pf^2 + n^{-1/2}\mathbb{G}_n f^2$ and $Pf^2 \le \delta^2 PF^2$ for every f, we have

$$\mathrm{E}_P^*\sigma_n^2 \le \delta^2\|F\|_{P,2}^2 + \frac{1}{\sqrt{n}}\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}^2}. \tag{2.3}$$

Here the empirical process in the second term can be replaced by the symmetrized empirical process $\mathbb{G}_n^o$ (defined as $\mathbb{G}_n^o f = n^{-1/2}\sum_{i=1}^n \varepsilon_i f(X_i)$ for independent Rademacher variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$) at the cost of adding a multiplicative factor 2 (e.g. [12], 2.3.1). The expectation can be factorized as the expectation on the Rademacher variables ε followed by the expectation on $X_1, \ldots, X_n$, and $\mathrm{E}_\varepsilon\|\mathbb{G}_n^o\|_{\mathcal{F}^2} \le 2\mathrm{E}_\varepsilon\|\mathbb{G}_n^o\|_{\mathcal{F}}$ by the contraction principle for Rademacher variables ([6], Theorem 4.12), and the fact that $F \le 1$ by assumption. Taking the expectation on $X_1, \ldots, X_n$, we obtain that $\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}^2} \le 4\mathrm{E}_P^*\|\mathbb{G}_n^o\|_{\mathcal{F}}$, which in turn is bounded above by $8\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}}$ by the desymmetrization inequality (e.g. 2.36 in [12]).

Thus $\mathcal{F}^2$ in the last term of (2.3) can be replaced by $\mathcal{F}$, at the cost of inserting a constant. Next we apply (2.2) to this term, and conclude that $z^2 := \mathrm{E}_P^*\sigma_n^2/\|F\|_{P,2}^2$ satisfies the inequality

$$z^2 \lesssim \delta^2 + \frac{J(z, \mathcal{F}, L_2)}{\sqrt{n}\|F\|_{P,2}}. \tag{2.4}$$

We apply Lemma 2.1 with $r = 1$, $A = \delta$ and $B^2 = 1/(\sqrt{n}\|F\|_{P,2})$ to see that

$$J(z, \mathcal{F}, L_2) \lesssim J(\delta, \mathcal{F}, L_2) + \frac{J^2(\delta, \mathcal{F}, L_2)}{\delta^2\sqrt{n}\|F\|_{P,2}}.$$

We insert this in (2.2) to complete the proof. □

Lemma 2.1. *Let $J\colon (0,\infty) \to \mathbb{R}$ be a concave, nondecreasing function with $J(0) = 0$. If $z^2 \le A^2 + B^2 J(z^r)$ for some $r \in (0,2)$ and $A, B > 0$, then*

$$J(z) \lesssim J(A)\left[1 + J(A^r)\left(\frac{B}{A}\right)^2\right]^{1/(2-r)}.$$

Proof. For $t > s > 0$ we can write s as the convex combination $s = (s/t)t + (1 - s/t)0$ of t and 0. Since $J(0) = 0$, the concavity of J gives that $J(s) \ge (s/t)J(t)$. Thus the function $t \mapsto J(t)/t$ is decreasing, which implies that $J(Ct) \le CJ(t)$ for $C \ge 1$ and any $t > 0$.

By the monotonicity of J and the assumption on z it follows that

$$J(z^r) \le J\Big(\big(A^2 + B^2 J(z^r)\big)^{r/2}\Big) \le J(A^r)\left(1 + \left(\frac{B}{A}\right)^2 J(z^r)\right)^{r/2}.$$

This implies that $J(z^r)$ is bounded by a multiple of the maximum of $J(A^r)$ and $J(A^r)(B/A)^r J(z^r)^{r/2}$. If it is bounded by the second one, then $J(z^r)^{1-r/2} \lesssim J(A^r)(B/A)^r$. We conclude that

$$J(z^r) \lesssim J(A^r) + J(A^r)^{2/(2-r)}\left(\frac{B}{A}\right)^{2r/(2-r)}.$$

Next again by the monotonicity of J,

$$\begin{aligned} J(z) &\le J\Big(\sqrt{A^2 + B^2 J(z^r)}\Big) \le J(A)\sqrt{1 + \left(\frac{B}{A}\right)^2 J(z^r)} \\ &\lesssim J(A)\left[1 + \left(\frac{B}{A}\right)^2\left(J(A^r) + J(A^r)^{2/(2-r)}\left(\frac{B}{A}\right)^{2r/(2-r)}\right)\right]^{1/2} \\ &\lesssim J(A)\left[1 + \sqrt{J(A^r)}\left(\frac{B}{A}\right) + \left(\frac{B}{A}\right)^{2/(2-r)} J(A^r)^{1/(2-r)}\right]. \end{aligned}$$

The middle term on the right side is bounded by a multiple of the sum of the first and third terms, since $x \lesssim 1^p + x^q$ for any conjugate pair (p, q) and any $x > 0$, in particular $x = \sqrt{J(A^r)}B/A$. □

For values of δ such that $\delta\|F\|_{P,2} \ll 1/\sqrt{n}$ Theorem 2.1 can be improved. (This seems not to be of prime interest for statistical applications.) Its bound can be written in the form $J(\delta, \mathcal{F}, L_2)\|F\|_{P,2} + J^2(\delta, \mathcal{F}, L_2)/(\delta^2\sqrt{n})$. In the second term δ can be replaced by $1/(\|F\|_{P,2}\sqrt{n})$, which is better if δ is smaller than the latter number, as the function $\delta \mapsto J(\delta, \mathcal{F}, L_2)/\delta$ is decreasing.

Lemma 2.2. *Under the conditions of Theorem 2.1,*

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\bigl(\delta,\mathcal{F},L_2\bigr)\|F\|_{P,2} + J^2\biggl(\frac{1}{\sqrt{n}\|F\|_{P,2}},\mathcal{F},L_2\biggr)\sqrt{n}\|F\|_{P,2}^2.$$

Proof. We follow the proof of Theorem 2.1 up to (2.4), but next use the alternative bounds

$$\begin{aligned} J(z,\mathcal{F},L_2) &\lesssim J\left(\sqrt{\delta^2 + \frac{J(z,\mathcal{F},L_2)}{\sqrt{n}\|F\|_{P,2}}},\mathcal{F},L_2\right)\\ &\le J(\delta,\mathcal{F},L_2) + J\left(\sqrt{\frac{J(z,\mathcal{F},L_2)}{\sqrt{n}\|F\|_{P,2}}},\mathcal{F},L_2\right)\\ &\le J(\delta,\mathcal{F},L_2) + J(\delta_n,\mathcal{F},L_2)\sqrt{\frac{J(z,\mathcal{F},L_2)}{\delta_n}\vee 1}, \end{aligned}$$

for $1/\delta_n = \sqrt{n}\|F\|_{P,2}$. Here we have used the subadditivity of the map $\delta \mapsto J(\delta,\mathcal{F},L_2)$, and the inequality $J(C\delta,\mathcal{F},l_2) \le CJ(\delta,\mathcal{F},L_2)$ for $C \ge 1$ in the last step. We can bound the sum of the three terms on the right side by a multiple of the maximum of these terms and conclude that the left side is smaller than at least one of the three terms. Solving next yields that

$$J(z,\mathcal{F},L_2) \lesssim J(\delta,\mathcal{F},L_2) \vee \frac{J^2(\delta_n,\mathcal{F},L_2)}{\delta_n} \vee J(\delta_n,\mathcal{F},L_2).$$

Because $J(\delta_n,\mathcal{F},L_2) \ge \delta_n$ for every $\delta_n > 0$, by the definition of the entropy integral, the third term on the right is bounded by the second term. We substitute the bound in (2.2) to finish the proof. □

3. Unbounded classes

In this section we investigate relaxations of the assumption that the class $\mathcal{F}$ of functions is uniformly bounded, made in Theorem 2.1. We start with a moment bound on the envelope.

Theorem 3.1. *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope function F such that $PF^{(4p-2)/(p-1)} < \infty$ for some $p > 1$ and such that $\mathcal{F}^2$ and $\mathcal{F}^4$ are P-measurable. If $Pf^2 < \delta^2 PF^2$ for every f and some $\delta \in (0,1)$, then*

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\bigl(\delta,\mathcal{F},L_2\bigr)\left(1 + \frac{J(\delta^{1/p},\mathcal{F},L_2)}{\delta^2\sqrt{n}}\frac{\|F\|_{P,(4p-2)/(p-1)}^{2-1/p}}{\|F\|_{P,2}^{2-1/p}}\right)^{p/(2p-1)}\|F\|_{P,2}.$$

Proof. Application of (2.1) to the functions f^2, forming the class $\mathcal{F}^2$ with envelope function F^2, yields

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}^2} \lesssim \mathrm{E}_P^* J\biggl(\frac{\sigma_{n,4}^2}{(\mathbb{P}_n F^4)^{1/2}},\mathcal{F}^2|\,F^2,L_2\biggr)\,(\mathbb{P}_n F^4)^{1/2}, \tag{3.1}$$

for $\sigma_{n,r}$ the diameter of $\mathcal{F}$ in $L_r(\mathbb{P}_n)$, i.e.

$$\sigma_{n,r}^r = \sup_f \mathbb{P}_n |f|^r. \tag{3.2}$$

Preservation properties of uniform entropy (see [10], or [12], 2.10.20, where the supremum over Q can also be moved outside the integral to match our current definition of entropy integral, applied to $\phi(f) = f^2$ with $L = 2F$) show that $J(\delta, \mathcal{F}^2 | F^2, L_2) \lesssim J(\delta, \mathcal{F} | F, L_2)$, for every $\delta > 0$. Because $\mathbb{P}_n f^2 = Pf^2 + n^{-1/2}\mathbb{G}_n f^2$ and $Pf^2 \leq \delta^2 PF^2$ by assumption, we find that

$$\mathrm{E}_P^* \sigma_{n,2}^2 \lesssim \delta^2 PF^2 + \frac{1}{\sqrt{n}} \mathrm{E}_P^* J\left(\frac{\sigma_{n,4}^2}{(\mathbb{P}_n F^4)^{1/2}}, \mathcal{F}, L_2 \right) (\mathbb{P}_n F^4)^{1/2}. \tag{3.3}$$

The next step is to bound $\sigma_{n,4}$ in terms of $\sigma_{n,2}$.

By Hölder's inequality, for any conjugate pair (p, q) and any $0 < s < 4$,

$$\mathbb{P}_n f^4 \leq \mathbb{P}_n |f|^{4-s} F^s \leq \left(\mathbb{P}_n |f|^{(4-s)p} \right)^{1/p} \left(\mathbb{P}_n F^{sq} \right)^{1/q}.$$

Choosing s such that $(4-s)p = 2$ (and hence $sq = (4p-2)/(p-1)$), we find that

$$\sigma_{n,4}^4 \leq \sigma_{n,2}^{2/p} \left(\mathbb{P}_n F^{sq} \right)^{1/q}.$$

We insert this bound in (3.3). The function $(x, y) \mapsto x^{1/p} y^{1/q}$ is concave, and hence the function $(x, y, z) \mapsto J(\sqrt{x^{1/p} y^{1/q}/z}, \mathcal{F}, L_2)\sqrt{z}$ can be seen to be concave by the same arguments as in the proof of Theorem 2.1. Therefore, we can apply Jensen's inequality to see that

$$\mathrm{E}_P^* \sigma_{n,2}^2 \lesssim \delta^2 PF^2 + \frac{1}{\sqrt{n}} J\left(\frac{(\mathrm{E}_P^* \sigma_{n,2}^2)^{1/(2p)} \left(PF^{sq} \right)^{1/(2q)}}{(PF^4)^{1/2}}, \mathcal{F}, L_2 \right) (PF^4)^{1/2}.$$

We conclude that $z := (\mathrm{E}_P^* \sigma_{n,2}^2)^{1/2} / \|F\|_{P,2}$ satisfies

$$\begin{aligned} z^2 &\lesssim \delta^2 + \frac{1}{\sqrt{n}} J\left(z^{1/p} \frac{(PF^2)^{1/(2p)} (PF^{sq})^{1/(2q)}}{(PF^4)^{1/2}}, \mathcal{F}, L_2 \right) \frac{(PF^4)^{1/2}}{PF^2} \\ &\lesssim \delta^2 + J(z^{1/p}, \mathcal{F}, L_2) \frac{(PF^{sq})^{1/(2q)}}{\sqrt{n}(PF^2)^{1-1/(2p)}}. \end{aligned}$$

In the last step we use that $J(C\delta, \mathcal{F}, L_2) \leq C J(\delta, \mathcal{F}, L_2)$ for $C \geq 1$, and Hölder's inequality as previously to see that the present C satisfies this condition. We next apply Lemma 2.1 (with $r = 1/p$) to obtain a bound on $J(z, \mathcal{F}, L_2)$, and conclude the proof by substituting this bound in (2.2). □

The preceding theorem assumes only a finite moment of the envelope function, but in comparison to Theorem 2.1 substitutes $J(\delta^{1/p}, \mathcal{F}, L_2)$ in the correction term of the upper bound, where $p > 1$ and hence $\delta^{1/p} \gg \delta$ for small δ. In applications to moduli of continuity of minimum contrast criteria this is

sufficient to obtain consistency with a rate, but typically the rate will be suboptimal. The rate improves as $p \downarrow 1$, which requires finite moments of the envelope function of order increasing to infinity, the limiting case $p = 1$ corresponding to a bounded envelope, as in Theorem 2.1. The following theorem interpolates between finite moments of any order and a bounded envelope function. If applied to obtaining rates of convergence it gives rates that are optimal up to a logarithmic factor.

Theorem 3.2. *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope function F such that $P\exp(F^{p+\rho}) < \infty$ for some $p, \rho > 0$ and such that $\mathcal{F}^2$ and $\mathcal{F}^4$ are P-measurable. If $Pf^2 < \delta^2 PF^2$ for every f and some $\delta \in (0, 1/2)$, then for a constant c depending on p, PF^2, PF^4 and $P\exp(F^{p+\rho})$,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \leq cJ\bigl(\delta, \mathcal{F}, L_2\bigr)\left(1 + \frac{J\bigl(\delta(\log(1/\delta))^{1/p}, \mathcal{F}, L_2\bigr)}{\delta^2\sqrt{n}}\right).$$

Proof. Fix $r = 2/p$. The functions $\psi, \overline{\psi}\colon [0,\infty) \to [0,\infty)$ defined by

$$\psi(f) = \log^r(1+f), \qquad \overline{\psi}(f) = e^{f^{1/r}} - 1,$$

are each other's inverses, and are increasing from $\psi(0) = \overline{\psi}(0) = 0$ to infinity. Thus their primitive functions $\Psi(f) = \int_0^f \psi(s)\,ds$ and $\overline{\Psi}(f) = \int_0^f \overline{\psi}(s)\,ds$ satisfy *Young's inequality* $fg \leq \Psi(f) + \overline{\Psi}(g)$, for every $f, g \geq 0$ (e.g. [2], page 120, 3.38).

The function $t \mapsto t\log^r(1/t)$ is concave in a neighbourhood of 0 (specifically: on the interval $(0, e^{1-r} \wedge 1)$), with limit from the right equal to 0 at 0, and derivative tending to infinity at this point. Therefore, there exists a *concave, increasing* function $k\colon (0,\infty) \to (0,\infty)$ that is identical to $t \mapsto t\log^r(1/t)$ near 0 and bounded below and above by a positive constant times the identity throughout its domain. (E.g. extend $t \mapsto t\log^r(1/t)$ linearly with slope 1 from the point where the derivative of the latter function has decreased to 1.) Write $k(t) = t\ell^r(t)$, so that ℓ^r is bounded below by a constant and $\ell(t) = \log(1/t)$ near 0. Then, for every $t > 0$,

$$\frac{\log(2 + t/C)}{\ell(C)} \lesssim \log(2+t). \tag{3.4}$$

(The constant in $\lesssim$ may depend on r.) To see this, note that for $C > c$ the left side is bounded by a multiple of $\log(2 + t/c)$, whereas for small C the left side is bounded by a multiple of $\bigl[\log(2+t) + \log(1 + 1/C)\bigr]/\ell(C) \lesssim \log(2+t) + 1$.

From the inequality $\Psi(f) \leq f\psi(f)$, we obtain that, for $f > 0$,

$$\Psi\left(\frac{f}{\log^r(2+f)}\right) \lesssim f.$$

Therefore, by (3.4) followed by Young's inequality,

$$\frac{f^4}{k(C^2)} = \frac{f^2/C^2}{\log^r(2+f^2/C^2)} \frac{f^2 \log^r(2+f^2/C^2)}{\ell^r(C^2)} \lesssim \frac{f^2}{C^2} + \overline{\Psi}\big(F^2 \log^r(2+F^2)\big).$$

On integrating this with respect to the empirical measure, with $C^2 = \mathbb{P}_n f^2$, we see that, with $G = \overline{\Psi}\big(F^2 \log^r(2+F^2)\big)$,

$$\mathbb{P}_n f^4 \lesssim k(\mathbb{P}_n f^2)\,\big(1 + \mathbb{P}_n G\big).$$

We take the supremum over f to bound $\sigma_{n,4}^4$ as in (3.2) in terms of $k(\sigma_{n,2}^2)$, and next substitute this bound in (3.3) to find that

$$\begin{aligned} \mathrm{E}_P^* \sigma_{n,2}^2 &\le \delta^2 PF^2 + \frac{1}{\sqrt{n}} \mathrm{E}_P^* J\left(\frac{\sqrt{k(\sigma_{n,2}^2)}\sqrt{1+\mathbb{P}_n G}}{(\mathbb{P}_n F^4)^{1/2}}, \mathcal{F}, L_2 \right) (\mathbb{P}_n F^4)^{1/2} \\ &\le \delta^2 PF^2 + \frac{1}{\sqrt{n}} J\left(\frac{\sqrt{k(\mathrm{E}_P^* \sigma_{n,2}^2)}\sqrt{1+PG}}{(PF^4)^{1/2}}, \mathcal{F}, L_2 \right) (PF^4)^{1/2}, \end{aligned}$$

where we have used the concavity of k, and the concavity of the other maps, as previously. By assumption the expected value PG is finite for $r = 2/p$. It follows that $z^2 = \mathrm{E}_P^* \sigma_{n,2}^2 / PF^2$ satisfies, for suitable constants a, b, c depending on r, PF^2, PF^4 and PG,

$$z^2 \lesssim \delta^2 + \frac{a}{\sqrt{n}} J(\sqrt{k(z^2 b)}c, \mathcal{F}, L_2).$$

By concavity and the fact that $k(0) = 0$, we have $k(Cz) \le Ck(z)$, for $C \ge 1$ and $z > 0$. The function $z \mapsto \sqrt{k(z^2 b)}c$ inherits this property. Therefore we can apply Lemma 3.1, with k of the lemma equal to the present function $z \mapsto \sqrt{k(z^2 b)}c$, to obtain a bound on $J(z, \mathcal{F}, L_2)$ in terms of $J(\delta, \mathcal{F}, L_2)$ and $J\big(\sqrt{k(\delta^2 b)}c, \mathcal{F}, L_2\big)$, which we substitute in (2.2). Here $k(\delta^2) = \delta^2 \log^r(1/\delta)$ for sufficiently small $\delta > 0$ and $\kappa(\delta^2) \lesssim \delta^2 \lesssim \delta^2 \log^r(1/\delta)$ for $\delta < 1/2$ and bounded away from 0. Thus we can simplify the bound to the one in the statement of the theorem, possibly after increasing the constants a, b, c to be at least 1, to complete the proof. □

Lemma 3.1. *Let $J\colon (0,\infty) \to \mathbb{R}$ be a concave, nondecreasing function with $J(0) = 0$, and let $k\colon (0,\infty) \to (0,\infty)$ be nondecreasing and satisfy $k(Cz) \le Ck(z)$ for $C \ge 1$ and $z > 0$. If $z^2 \le A^2 + B^2 J\big(k(z)\big)$ for some $A, B > 0$, then*

$$J(z) \lesssim J(A)\left[1 + J\big(k(A)\big)\left(\frac{B}{A}\right)^2\right].$$

Proof. As noted in the proof of Lemma 2.1 the properties of J imply that $J(Cz) \le CJ(z)$ for $C \ge 1$ and any $z > 0$. In view of the assumed property of k and the monotonicity of J it follows that $J \circ k(Cz) \le CJ \circ k(z)$ for every $C \ge 1$ and $z > 0$. Therefore, by the monotonicity of J and k, and the assumption on z,

$$J \circ k(z) \le J \circ k\Big(\sqrt{A^2 + B^2 J \circ k(z)}\Big) \le J \circ k(A)\sqrt{1 + (B/A)^2 J \circ k(z)}.$$

As in the proof of Lemma 2.1 we can solve this for $J \circ k(z)$ to find that

$$J \circ k(z) \lesssim J \circ k(A) + J \circ k(A)^2 \left(\frac{B}{A}\right)^2.$$

Next again by the monotonicity of J,

$$\begin{aligned} J(z) &\le J\Big(\sqrt{A^2 + B^2 J \circ k(z)}\Big) \le J(A)\sqrt{1 + (B/A)^2 J \circ k(z)} \\ &\lesssim J(A)\left[1 + \left(\frac{B}{A}\right)\sqrt{J \circ k(A)} + \left(\frac{B}{A}\right)^2 J \circ k(A)\right]. \end{aligned}$$

The middle term on the right side is bounded by the sum of the first and third terms. □

References

[1] Birgé, L., and Massart, P. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields 97*, 1-2 (1993), 113–150. MR1240719

[2] Boyd, S., and Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge, 2004. MR2061575

[3] Dudley, R. M. Central limit theorems for empirical measures. *Ann. Probab. 6*, 6 (1978), 899–929 (1979). MR0512411

[4] Giné, E., and Koltchinskii, V. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab. 34*, 3 (2006), 1143–1216. MR2243881

[5] Kolchins'kiĭ, V. Ī. On the central limit theorem for empirical measures. *Teor. Veroyatnost. i Mat. Statist. 24* (1981), 63–75, 152. MR0628431

[6] Ledoux, M., and Talagrand, M. *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes. MR1102015

[7] Massart, P., and Nédélec, É. Risk bounds for statistical learning. *Ann. Statist. 34*, 5 (2006), 2326–2366. MR2291502

[8] Ossiander, M. A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab. 15*, 3 (1987), 897–919. MR0893905

[9] Pollard, D. A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A 33*, 2 (1982), 235–248. MR0668445

[10] Pollard, D. *Empirical processes: theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990. MR1089429

[11] van de Geer, S. The method of sieves and minimum contrast estimators. *Math. Methods Statist. 4*, 1 (1995), 20–38. MR1324688

[12] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. MR1385671